Lyve-SET: an hqSNP pipeline for outbreak investigations

Lee Katz, Ph.D.

Bioinformatics scientist

Enteric Diseases Laboratory Branch (EDLB)

Centers for Disease Control and Prevention (CDC)

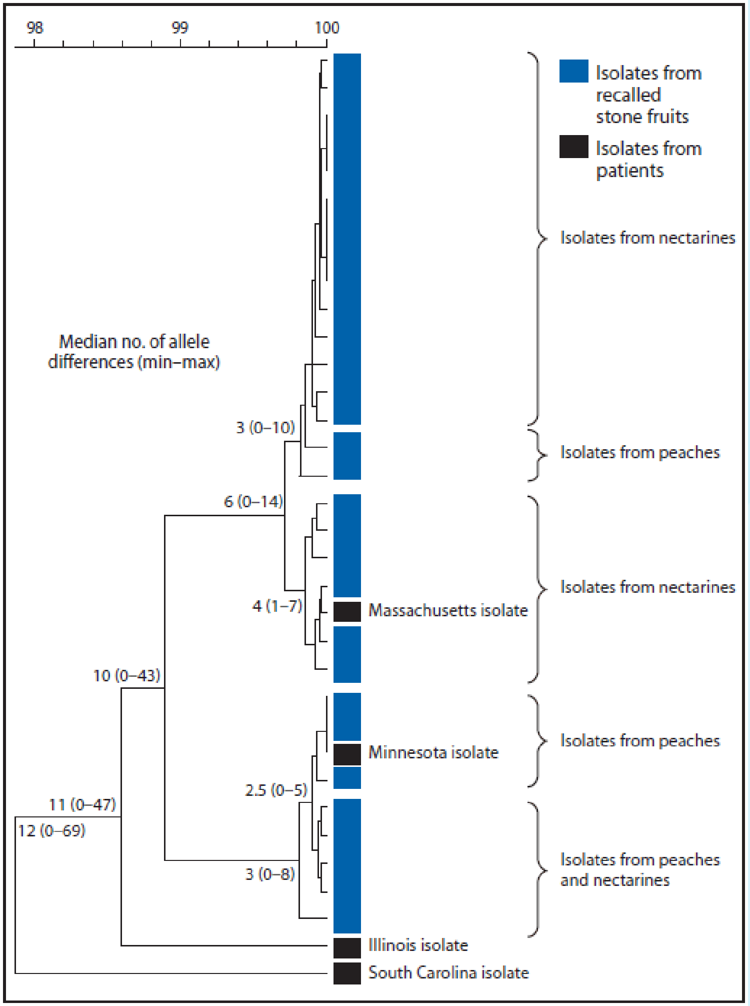
Sequencing, Finishing, and Analysis of the Future 2015

May 28, 2015

Enteric Diseases Laboratory Branch (EDLB) at CDC

- We study bacterial foodborne pathogens: *Listeria monocytogenes*, *E. coli*, *Salmonella*, *C. botulinum*,…
- Perform routine surveillance

Traditionally, tracked by PFGE (and other methods, e.g., 7-gene MLST)



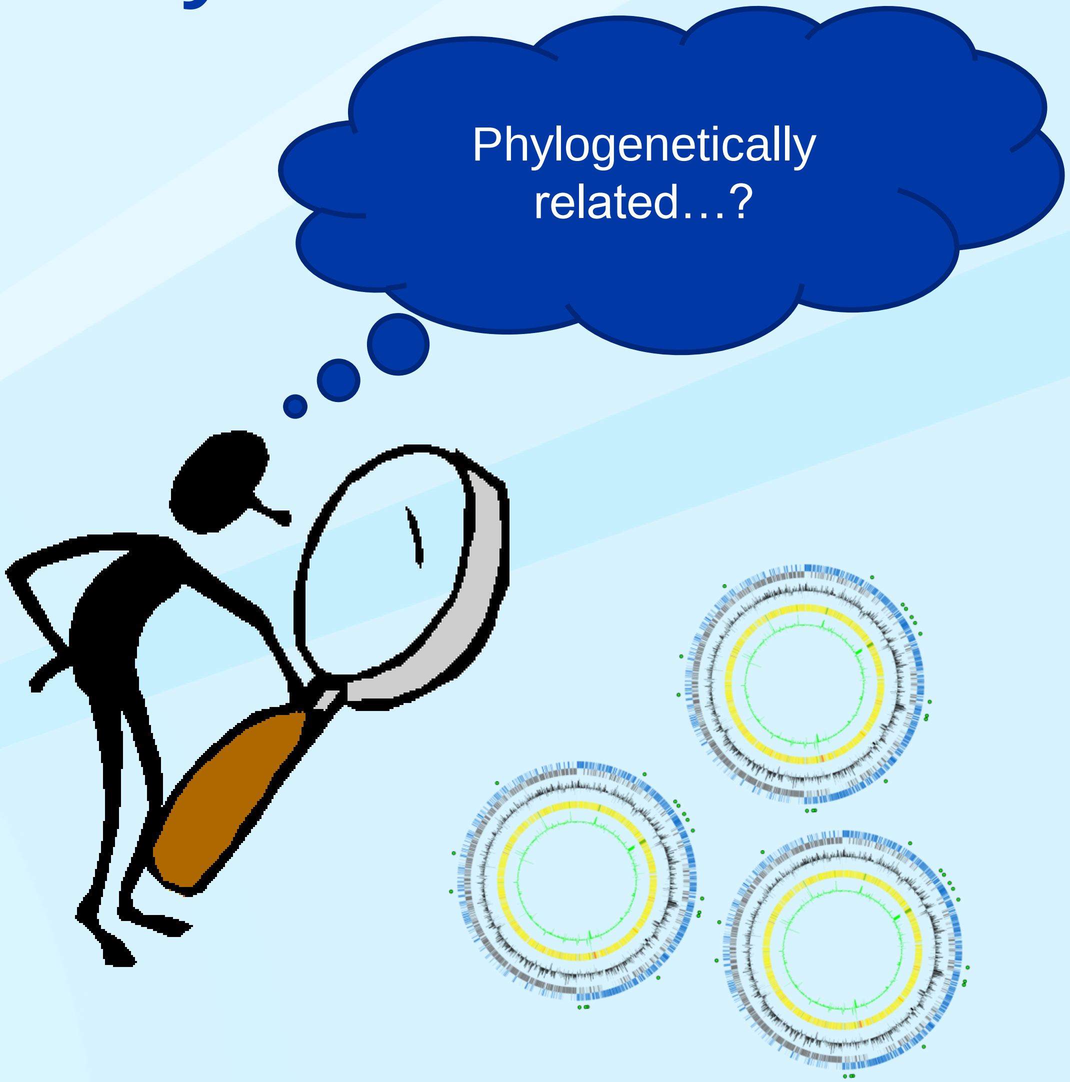
But now, transitioning to whole-genome methods for routine surveillance

PFGE photo taken from http://www.cdc.gov/pulsenet/about/faq.html

Phylogeny taken from Jackson et al 2015, “Notes from the Field: Listeriosis Associated with Stone Fruit — United States, 2014” MMWR

We need high-quality SNPs!

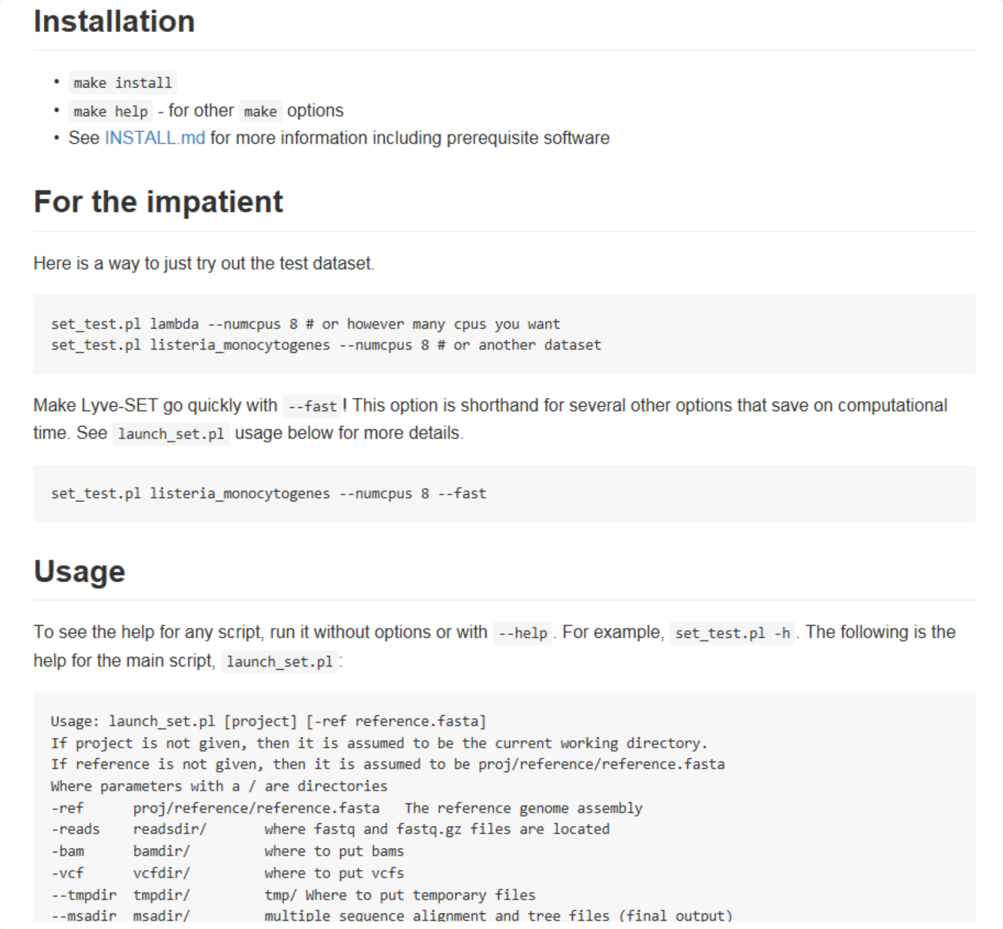
- High-quality SNPs (hqSNPs) can give a fine-resolution view of a cluster of genomes
- Useful for outbreak investigations
- Therefore, we created Lyve-SET for hqSNP-based phylogenies



Lyve-SET

Lyve – *Listeria, Yersinia, Vibrio,* and *Enterobacteriaceae* reference lab
SET – Snp Extraction Tool

- Some details on Lyve-SET:
  - For LINUX
  - Extensive documentation
  - Help options are embedded in each script
  - `--fast` option, takes ¼ the normal time
  - Easy to use
  - Modular
- Used in labs at CDC: Foodborne Diseases Laboratory Branch, Clinical and Environmental Microbiology Branch, Respiratory Diseases Branch
- *Listeria monocytogenes* outbreak investigations since summer 2013



Installation

- `make install`
- `make help` - for other `make` options
- See INSTALL.md for more information including prerequisite software

For the impatient

Here is a way to just try out the test dataset.

```
set_test.pl lambda --numcpus 8 # or however many cpus you want
set_test.pl listeria_monocytogenes --numcpus 8 # or another dataset
```

Make Lyve-SET go quickly with `--fast` ! This option is shorthand for several other options that save on computational time. See `launch_set.pl` usage below for more details.

```
set_test.pl listeria_monocytogenes --numcpus 8 --fast
```

Usage

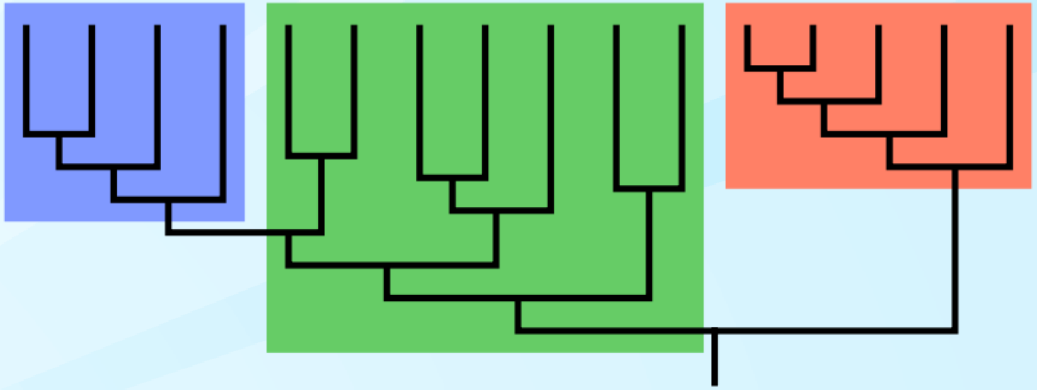
To see the help for any script, run it without options or with `--help` . For example, `set_test.pl -h` . The following is the help for the main script, `launch_set.pl` :

```
Usage: launch_set.pl [project] [-ref reference.fasta]
If project is not given, then it is assumed to be the current working directory.
If reference is not given, then it is assumed to be proj/reference/reference.fasta
Where parameters with a / are directories
-ref      proj/reference/reference.fasta   The reference genome assembly
-reads    readsdir/       where fastq and fastq.gz files are located
-bam      bamdir/         where to put bams
-vcf      vcfdir/         where to put vcfs
--tmpdir  tmpdir/         tmp/ Where to put temporary files
--msadir  msadir/         multiple sequence alignment and tree files (final output)
```

https://github.com/lskatz/lyve-SET

HIGH-QUALITY SNPS: ASSUMPTIONS IN OUTBREAK INVESTIGATIONS

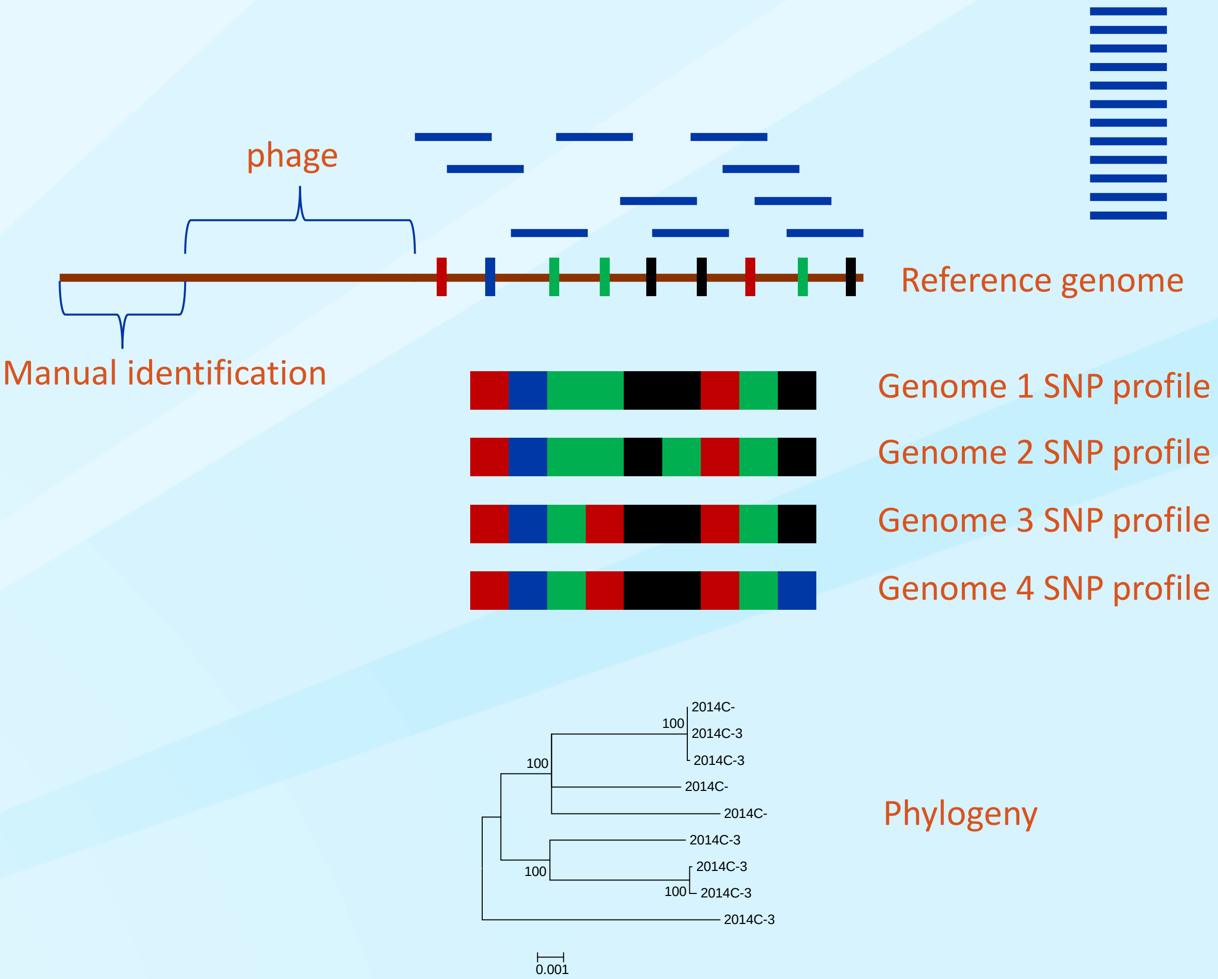
1. Evolution approximates epidemiology
2. SNPs correlate well with overall evolutionary change



A survey of SNP vs recombination rates in bacteria: Vos M, Didelot X. 2008. A comparison of homologous recombination rates in bacteria and archaea. ISME J 3:199-208.

An animation of Lyve-SET (hqSNPs)

0. Pre-processing
   - a) phage discovery/masking
   - b) Manual identification of troublesome regions
   - c) Read cleaning (Poster – Wagner et al)
1. Mapping - SMALT
   - a) 95% read identity
   - b) Unambiguous mapping
2. SNP calling - VarScan
   - a) 75% consensus
   - b) 10x depth
3. Phylogeny inferring – RAxML v8
   - a) Removal of clustered SNPs
   - b) Ascertainment bias model
   - c) Maximum likelihood



https://github.com/lskatz/lyve-SET/blob/master/docs/FAQ.md

https://www.sanger.ac.uk/resources/software/smalt

Koboldt, D., Zhang, Q., Larson, D., Shen, D., McLellan, M., Lin, L., Miller, C., Mardis, E., Ding, L., & Wilson, R. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing *Genome Research* DOI: 10.1101/gr.129684.111

Stamatakis, A. (2014) RAxML version8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. doi:10.1093/bioinformatics/btu033

Comparing against other well regarded tools

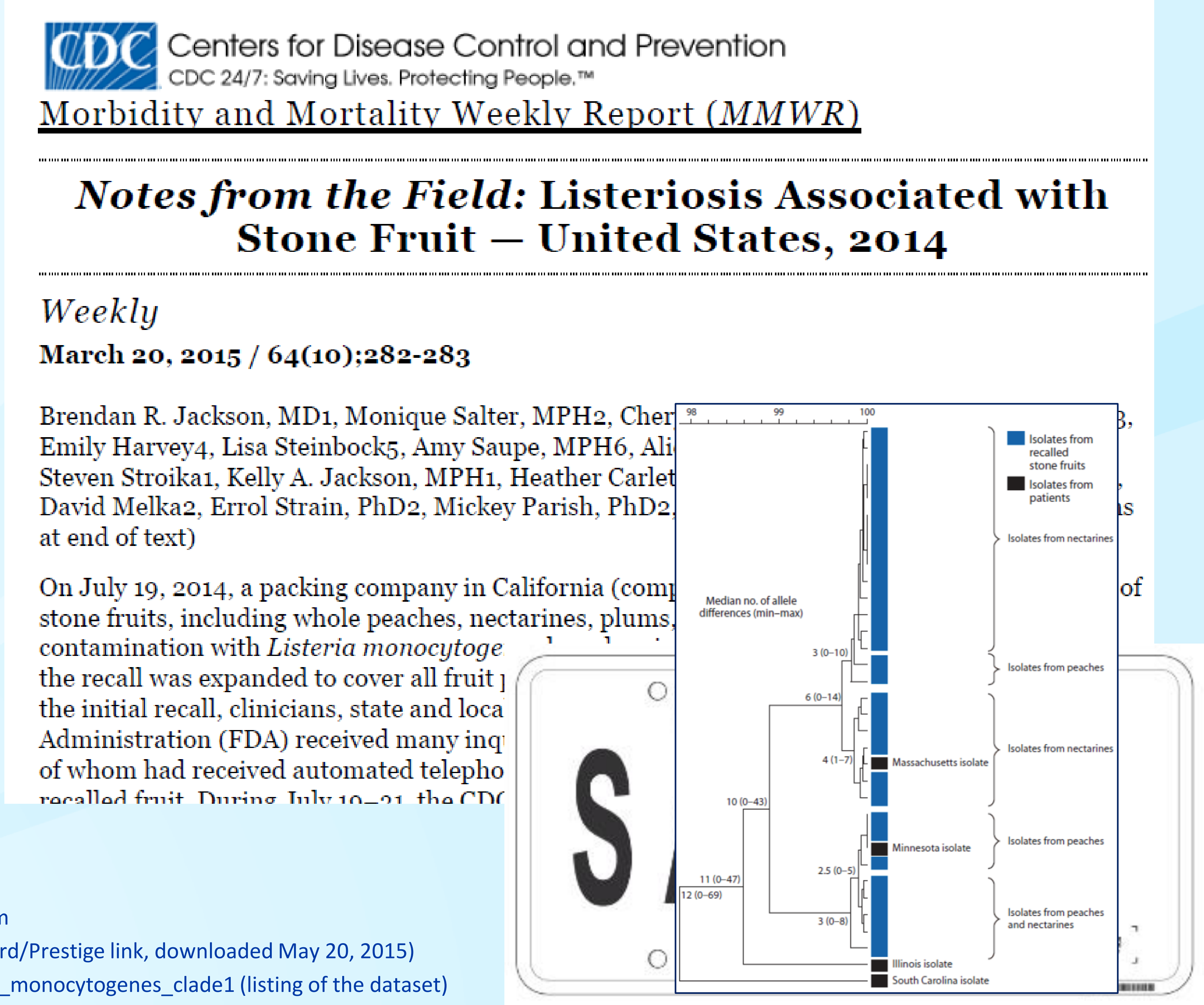
- **wgMLST**
  - Applied Maths: http://www.applied-maths.com/applications/wgmlst
  - International Listeria wgMLST Schema Development Consortium
- **SNVPhyl**
  - Petkau A, Keddy A, Slusky L, Mabon P, Bristow F, Matthews T, Adam J, Carriço JA, Katz LS, Reimer A, Knox N, Courtot M, Graham M, Hsiao W, Brinkman F, Beiko RG, Van Domselaar G. Outbreak investigation with IRIDA's SNVPhyl pipeline and GenGIS. Poster presented at: The 7th Meeting of the Global Microbial Identifier; September 11-12, 2014; York, UK
  - https://github.com/apetkau/core-phylogenomics
- **Snp-Pipeline v3.3**
  - Pettengill JB, Luo Y, Davis S, Chen Y, Gonzalez-Escalona N, Ottesen A, Rand H, Allard MW, Strain E An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: A case study with Salmonella.
  - http://snp-pipeline.readthedocs.org/en/latest
- **kSNP2**
  - Gardner, S.N. and Hall, B.G. 2013. When whole-genome alignments just won't work: kSNP v2 software for alignment-free SNP discovery and phylogenetics of hundreds of microbial genomes. PLoS ONE, 8(12):e81760.doi:10.1371/journal.pone.0081760
- **Wombac v2.1**
  - https://github.com/tseemann/wombac

Quick comparison with well-regarded tools

	Lyve-SET	*--fast*	wgMLST	SNVPhyl	Snp-Pipeline	kSNP	Wombac
Phage masking	x		x				
Manual masking	x			x			
Read cleaning	x		x	x	x		
HPC support	SGE	SGE	SGE	SGE, Torque, etc	SGE, Torque		
Customizable thresholds	x	x		x	x	x	x
Considers clustered SNPs	x	x	x			x	
Finished product	ML tree	ML tree	UPGMA, alleles	ML tree	SNP matrix	ML, NJ tree	ML tree
Approach	Ref-mapping	Ref-mapping	MLST-mapping	Ref-mapping	Ref-mapping	Asm-free	Ref-mapping
O/S	🐧	🐧	🪟+🐧	🍎🪟🐧	🐧	🐧	🐧
Availability	Open	Open	©	Open	Open	Open	Open

Stone Fruit outbreak/*Listeria*

- Summer 2014
- Contaminated stone fruit – peaches, nectarines, etc
- Two confirmed clinical cases, two related but sporadic cases, many environmental isolates
- Very good epidemiology; well characterized



Centers for Disease Control and Prevention
CDC 24/7: Saving Lives. Protecting People.™

Morbidity and Mortality Weekly Report (*MMWR*)

Notes from the Field: Listeriosis Associated with Stone Fruit — United States, 2014

Weekly

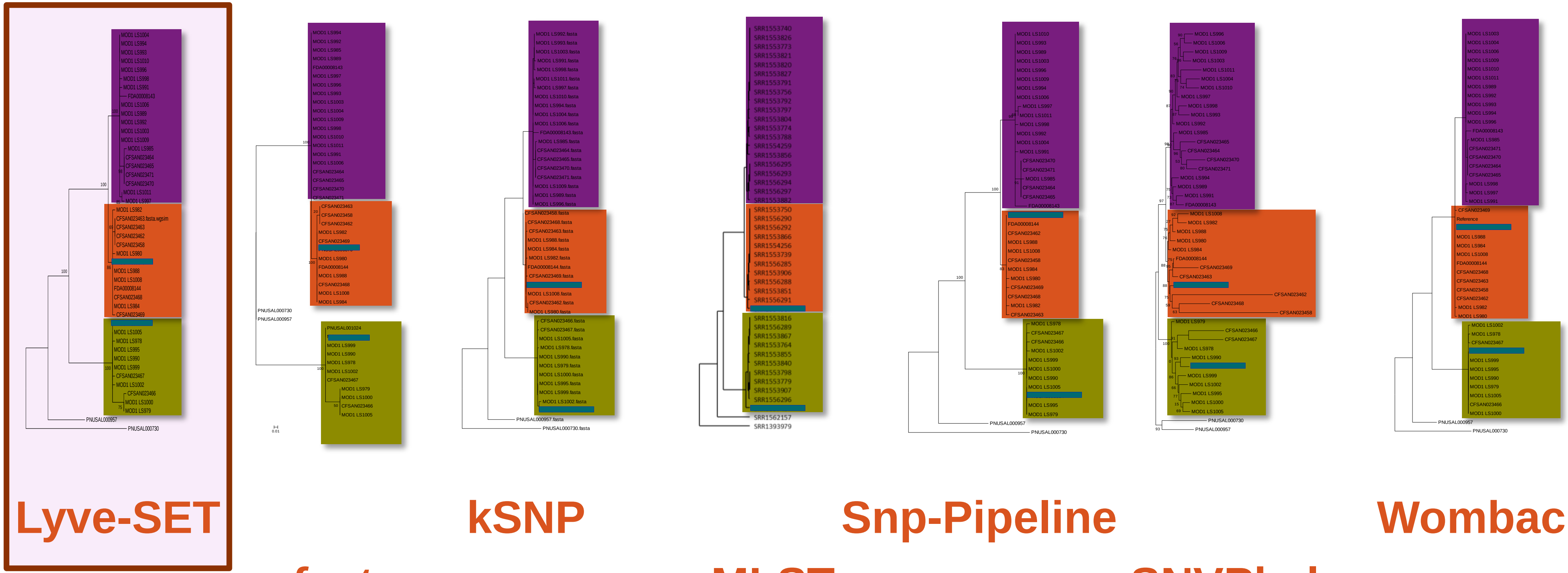
March 20, 2015 / 64(10);282-283

Brendan R. Jackson, MD1, Monique Salter, MPH2, Cher 3, Emily Harvey4, Lisa Steinbock5, Amy Saupe, MPH6, Ali Steven Stroika1, Kelly A. Jackson, MPH1, Heather Carlet , David Melka2, Errol Strain, PhD2, Mickey Parish, PhD2, ns at end of text)

On July 19, 2014, a packing company in California (comp of stone fruits, including whole peaches, nectarines, plums, contamination with *Listeria monocytoge* the recall was expanded to cover all fruit p the initial recall, clinicians, state and loca Administration (FDA) received many inq of whom had received automated telepho recalled fruit. During July 19–21, the CDC

http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6410a6.htm
http://mvd.dor.ga.gov/motor/plates/PlateSelection.aspx (Standard/Prestige link, downloaded May 20, 2015)
https://github.com/lskatz/lyve-SET/tree/master/testdata/listeria_monocytogenes_clade1 (listing of the dataset)

Comparison of the *Listeria monocytogenes* stone fruit outbreak trees.



Lyve-SET

--fast

kSNP

wgMLST

Snp-Pipeline

SNVPhyl

Wombac

- Both outbreak genomes cluster with the correct clades with 100% in all trees
- Most trees have almost the exact same topology with high confidence values for outbreak clades

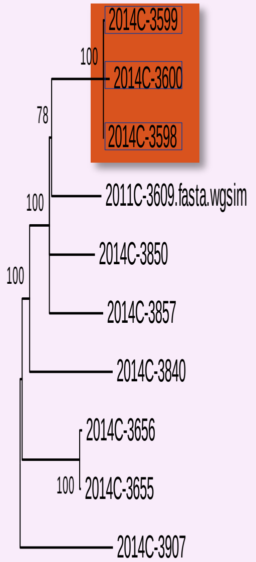
Sprouts/E. coli

- 2014
- 19 cases
- Raw clover sprouts
- Very good epidemiology; well characterized

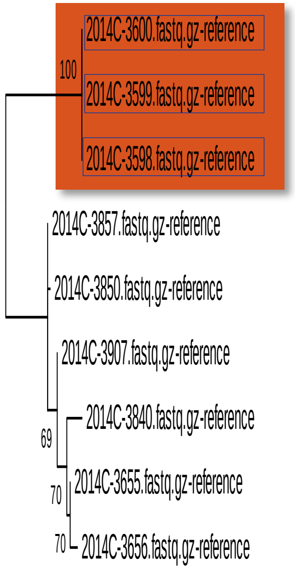
http://www.cdc.gov/ecoli/2014/O121-05-14/index.html

https://github.com/lskatz/lyve-SET/tree/master/testdata/escherichia_coli (listing of the dataset)

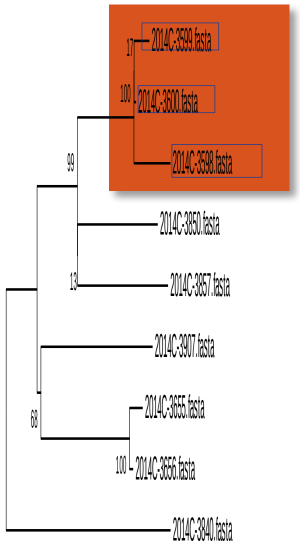
Comparison of the *E. coli sprouts* outbreak trees.



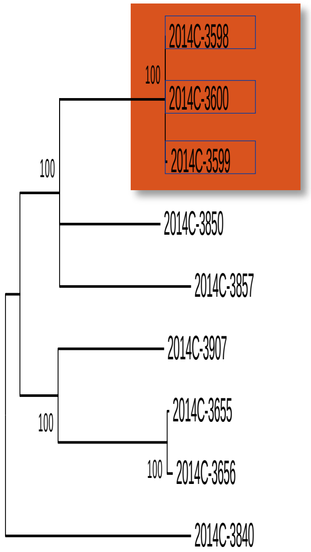
Lyve-SET


2014C-3600.fastq.gz-reference
100
2014C-3599.fastq.gz-reference
2014C-3598.fastq.gz-reference
2014C-3857.fastq.gz-reference
2014C-3850.fastq.gz-reference
2014C-3907.fastq.gz-reference
2014C-3840.fastq.gz-reference
69
2014C-3655.fastq.gz-reference
70
70
2014C-3656.fastq.gz-reference

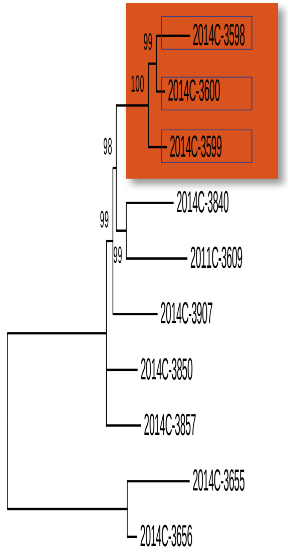
--fast



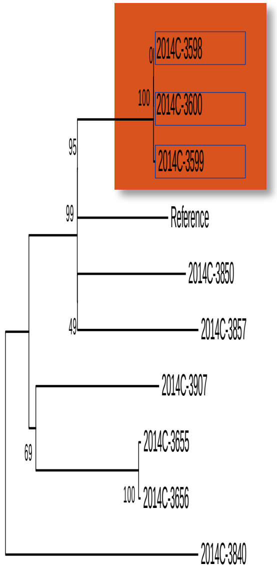
kSNP



Snp-Pipeline



SNVPhyl



Wombac

- Both outbreak genomes cluster with the correct clade with 100% in all trees
- Most trees have almost the exact same topology with high confidence values for outbreak clades

Other advantages of Lyve-SET

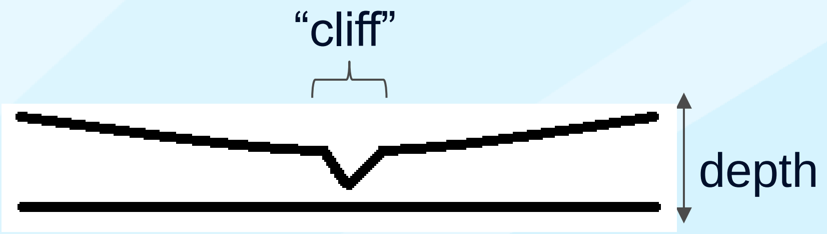
- Closely developed alongside outbreak investigations
- Modular – UNIX philosophy: each script does one thing and does it well.
  - Can switch in and out new scripts as desired
- Integration with CG-Pipeline (downsampling, read-cleaning, read-metrics, etc)
- Easy-to-understand documentation
- Easy to install
- Users mailing list
- Actively maintained

Thank you to the WGS standards and analysis working group for their contributions on the standardized datasets

https://github.com/lskatz/CG-Pipeline

Future work

- Avoiding SNP noise
  - Cliff detection
  - Soft-clipping of reads
- Annotation of SNPs
- Validation of SNPs with WGS standards/analysis working group
  - Members from FDA, USDA, NCBI, and CDC
  - Manually validate less-confident SNP calls
  - Create gold standard datasets



All proposed improvements: https://github.com/lskatz/lyve-SET/issues

http://www.taverna.org.uk

https://github.com/ssadedin/bpipe

https://github.com/lskatz/lyve-SET

Conclusions

- Aids in epidemiological investigations
- Gives concordant results
- Epidemiologically focused

https://groups.google.com/forum/#!forum/lyve-set

Thank you!

Authors: **Lee S. Katz**, Darlene D. Wagner, Aaron Petkau, Cameron Sieffert, Heather Carleton, Shaun Tyler, Gary Van Domselaar

- WGS standards working group
- Many others in EDLB/CDC
- Bioinformatics core at PHAC
- My wife

Lkatz@cdc.gov